УДК 004.8

Н.А. Новоселова, И.Э. Том
Объединенный институт проблем информатики
Национальной академии наук Беларуси, г. Минск
novosel@newman.bas-net.by, tom@newman.bas-net.by

Подход к построению ансамбля классификаторов с использованием генетического алгоритма

В статье рассматривается новый эволюционный подход к построению ансамбля классификаторов. Предложенный подход разработан на основе генетического алгоритма с модифицированной схемой реализации. В процессе оптимизации происходит определение параметров как отдельных классификаторов, так и всего ансамбля. С использованием подхода выполнено построение ансамбля классификаторов на нескольких наборах данных из архива данных по машинному обучению и на одном реальном наборе медицинских данных. Сравнительное тестирование показало преимущества использования предложенного подхода при работе с многомерными данными, характеризующимися большим количеством признаков.

Введение

Согласно литературным источникам [1], [2] использование комбинации классификаторов позволяет повысить точность классификации при решении практических задач. Среди всех имеющихся методов построения ансамбля классификаторов наиболее популярными являются «bagging» и «boosting» [3], которые основаны на манипуляциях с исходным обучающим множеством с целью построения нескольких классификаторов. Теоретические и эмпирические результаты показывают, что результат комбинации классификаторов наиболее эффективен, когда классификаторы являются независимыми [4]. Для построения независимых классификаторов наиболее эффективным методом является обучение отдельных членов ансамбля на различающихся подмножествах признаков [5], [6]. Таким образом, построение ансамбля классификаторов на основе декомпозиции исходного набора признаков, описывающих объекты данных, в большинстве случаев имеет преимущества. Известно большое количество публикаций, исследующих свойства ансамблей классификаторов, которые построены с использованием различных подмножеств признаков. Например, в работе [5] была продемонстрирована возможность использования рандомизированных подмножеств признаков для построения ансамбля классификаторов. Однако, когда размерность признакового пространства достаточно большая, такой способ является неэффективным. В работе [2] использовался эвристический алгоритм для декомпозиции множества признаков на несколько некоррелированных подмножеств, который, являясь локально оптимальным, не гарантировал получение наилучшего результата.

В настоящей работе представлен подход к построению ансамбля классификаторов, отличительной особенностью которого является использование генетического алгоритма (ГА) для одновременного отбора нескольких подмножеств признаков для построения отдельных классификаторов, входящих в состав ансамбля. Использова-

ние ГА для решения оптимизационной задачи декомпозиции исходного множества признаков для построения ансамбля классификаторов объясняется следующими причинами:

– простотой кодирования решения оптимизационной задачи;

– отсутствием ограничений на гладкость оптимизируемой функции, что позволяет в качестве последней использовать точность классификации с использованием ансамбля;

– отсутствием эффективных субоптимальных алгоритмов отбора подмножеств признаков для классификаторов, составляющих ансамбль.

В предыдущих работах [7-9] ГА использовался в основном для оптимизации отбора информативных признаков для построения индивидуального классификатора. В этом случае все множество признаков разбивалось на два подмножества, одно из которых полностью отбрасывалось и не использовалось при решении классификационной задачи. Предлагаемый в настоящей статье подход позволяет использовать все исходное множество признаков для построения ансамбля классификаторов с одновременным обучением как параметров индивидуальных классификаторов, так и всего ансамбля.

1. Формальное определение ансамбля классификаторов

Пусть имеется множество $\Omega = \{\omega_1, \ldots, \omega_c\}$ меток классов и пусть $x = [x_1, \ldots, x_M]^T \in$ $\in R^M$ – набор признаков, описывающих объект данных. Классификатором является отображение следующего вида:

$$D : R^M \to [0,1]^c,$$

где $D(x)$ – вектор размерности c, у которого i-й компонент определяет степень принадлежности x классу ω_i, $i = 1,\ldots,c$. В системах, основанных на комбинации k классификаторов, выходы отдельных классификаторов агрегируются для получения окончательного классификационного решения:

$$D(x) = F(D_1(x), \ldots, D_k(x)),$$

где F – оператор агрегирования. Выходом каждого отдельного классификатора для некоторого объекта данных x является c-мерный вектор $D_i(x) = [d_{i,1}(x), \ldots, d_{i,c}(x)]^T$, $i = 1,\ldots,k$. Выходом всей комбинации классификаторов является c-мерный вектор – $D(x) = [\mu_1(x), \ldots, \mu_c(x)]^T$. Если необходимо определить для объекта x единственную метку класса, то класс ω_s соответствует максимальному значению степеней принадлежности:

$$d_{i,s}(x) \ge d_{i,j}(x)\ \forall j = 1,\ldots,c \text{ – для отдельных классификаторов;}$$

$$\mu_s(x) \ge \mu_t(x),\ \forall t = 1,\ldots,c \text{ – для всего ансамбля.}$$

Существуют различные операторы, позволяющие комбинировать выходы отдельных классификаторов ансамбля. К ним относятся: оператор максимума, минимума, произведения, усреднения, решение «большинством голосов» и т.д. В нашем исследовании отдельные классификаторы комбинируются с использованием метода «большинством голосов», который является достаточно популярным и простым в реализации.

Пусть c-мерный вектор $D_i(x) = [d_{i,1}(x), \ldots, d_{i,c}(x)]^T \in [0,1]^c$ – выход классификатора D_i, i = *1*,…,k для входного объекта x. Значение $d_{i,j}(x) \in [0,1]$ означает поддержку,

оказываемую классификатором D_i гипотезе о том, что x относится к классу ω_j. Для того, чтобы определить «голос» классификатора в поддержку единственного класса, мы огрубляем классификационное решение, а именно выбираем класс

$$\omega_s \Leftrightarrow d_{i,s}(x) = \max_j \left\{ d_{i,j}(x) \right\}.$$

Таким образом, классификационное решение для каждого D_i формулируется как бинарный вектор D_i^h, имеющий единицу в позиции s и ноль в остальных позициях:

$$d_{i,j}^h(x) = \begin{cases} 1, & j = s \\ 0, & j \neq s \end{cases}.$$

Решение «большинством голосов» F_{maj}, представленное в виде c-мерного вектора, рассчитывается следующим образом:

$$F_{maj} \equiv D(x) = \left[d_1(x), \ldots, d_c(x) \right]^T, \quad d_j(x) \in \{0,1\}, j = 1,\ldots,c$$

и

$$d_j(x) = \begin{cases} 1, & \sum_{i=1}^{k} d_{i,j}^h(x) = \max_{s=1,\ldots,c} \sum_{i=1}^{k} d_{i,s}^h(x) \\ & 0, \end{cases},$$

где k – количество классификаторов в ансамбле.

В нашем исследовании мы используем различные подмножества исходных признаков для построения ансамбля классификаторов. В качестве отдельного классификатора используется метод ближайших соседей [10].

2. Подход к построению ансамбля классификаторов

Предложенный подход к построению ансамбля классификаторов разработан на основе ГА, который имеет модифицированную схему реализации применительно к задаче оптимизации разбиения множества признаков на подмножества, определяющие отдельные классификаторы ансамбля. Таким образом, формулируется следующая оптимизационная задача:

Пусть Ф – множество различных разбиений множества признаков, характеризующих объект данных, на k подмножеств, каждое из которых соответствует отдельному классификатору. Каждое разбиение представляет собой некоторую комбинацию входных признаков из максимально возможного количества комбинаций $(k+1)^M$, где M – количество входных признаков. Требуется найти такое разбиение $S \in \Phi$, которое является решением задачи оптимизации с одним критерием:

$$\max f_1(S),$$

где $f_1(S)$ – количество правильно классифицированных объектов с использованием ансамбля классификаторов.

Общая схема реализации предложенного подхода с использованием ГА представлена на рис. 1. Согласно рис. 1, случайным образом формируется поколение ГА путем различных разбиений всего множества признаков A обучающей выборки на k подмножеств A^j, $1 \leq j \leq k$. С использованием каждого из подмножеств признаков, закодированных в k отдельных особях ГА, выполняется построение k классификаторов. Классификационные решения отдельных классификаторов комбинируются с использованием рассмотренного выше оператора агрегирования «большинством голосов», определяя решение ансамбля классификаторов. Затем в цикле выполняются генетические операции рекомбинации и отбора особей ГА в новое поколение решений оптимизационной задачи, где в качестве функции приспособленности особи выступает результат классификации данных ансамблем классификаторов.

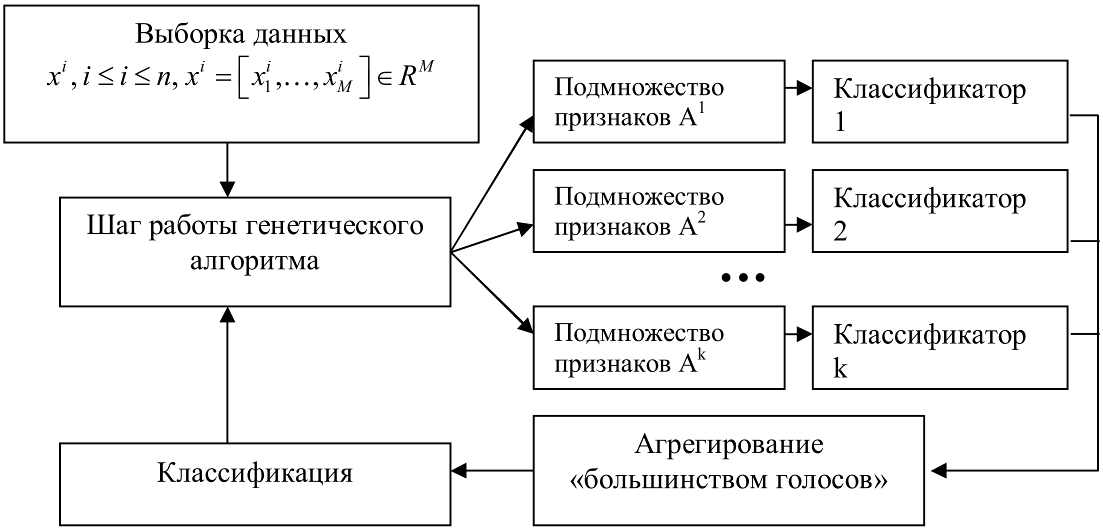


Рисунок 1 – Схема предложенного подхода

Одним из ключевых вопросов, возникающих при использовании ГА для решения прикладной задачи, является способ кодирования решения в особи, подвергающейся воздействию генетических операторов. В нашем исследовании в особи (хромосоме) ГА кодируется решение задачи разбиения множества признаков на подмножества для построения ансамбля классификаторов. Особь представляет собой множество признаков, каждый из которых отнесен к некоторому подмножеству, i-й ген соответствует i-му признаку. Были использованы две схемы кодирования:

1. В первой схеме каждый ген принимает значение от 1 до k, которое соответствует подмножеству признаков, определяющему индивидуальный классификатор. В этом случае множество исходных признаков делится на несколько непересекающихся подмножеств. Пространство поиска равно $(k+1)^M$, где M – количество входных признаков, например, при $k = 3$, и количеству признаков $M = 7$, возможное представление особи ГА представлено на рис. 2.

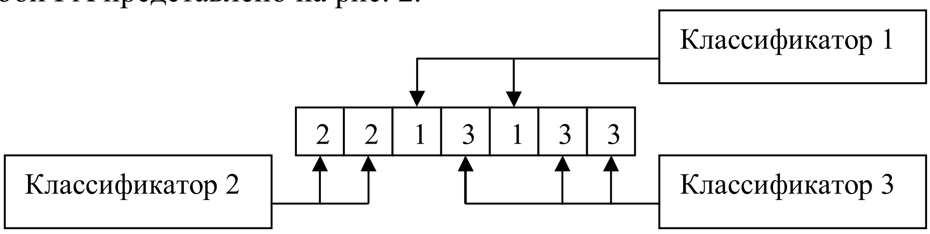


Рисунок 2 – Первая схема кодирования особи ГА

2. Во второй схеме существует возможность определения пересекающихся подмножеств признаков. Размерность пространства поиска равна $(2^k)^M$, где M – количество входных признаков. Пример кодирования особи ГА с тремя классификаторами и количеством признаков $M = 7$ представлен на рис. 3.

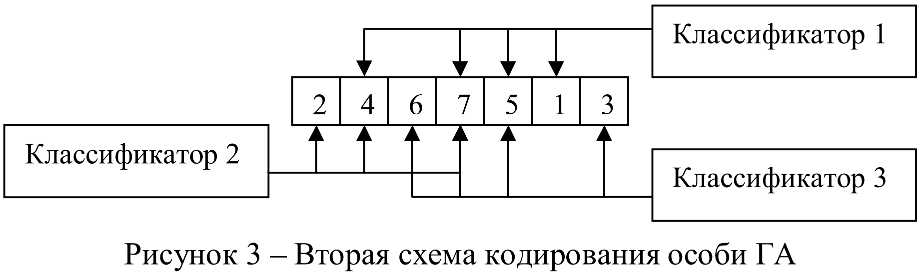


Рисунок 3 – Вторая схема кодирования особи ГА

Для кодирования представленной на рис. 3 особи используются следующие обозначения:
если значение гена равно 1 – признак принадлежит только первому подмножеству;
если значение гена равно 2 – признак принадлежит только второму подмножеству;
если значение гена равно 3 – признак принадлежит только третьему подмножеству;
если значение гена равно 4 – признак принадлежит только первому и второму подмножеству;
если значение гена равно 5 – признак принадлежит только первому и третьему подмножеству;
если значение гена равно 6 – признак принадлежит только второму и третьему подмножеству;
если значение гена равно 7 – признак принадлежит одновременно всем трем подмножествам.

3. Результаты экспериментов

Разработанный подход к построению ансамбля классификаторов с использованием ГА был протестирован (табл. 1) на двух наборах данных из архива данных по машинному обучению (http://www.ics.uci.edu/~mlearn/): по болезни сердца Heart, по определению типов вин Wine, и на одном наборе медицинских данных пациентов с транзиторными ишемическими атаками (ТИА)*.

Для оценки точности классификации ансамблем классификаторов мы разбивали исследуемые наборы данных на две части, одна из которых использовалась для обучения ансамбля (обучающая выборка), а вторая – для тестирования результатов (тестовая выборка).

Таблица 1 – Описание наборов данных для тестирования

Набор данных	Количество объектов данных	Количество признаков	Количество классов
Heart	303	13	2
Wine	178	13	3
ТИА	101	41	4

Было поставлено четыре эксперимента по построению ансамбля классификаторов:

1) построение ансамбля из трех классификаторов на основе непересекающихся подмножеств признаков;
2) построение ансамбля из пяти классификаторов на основе непересекающихся подмножеств признаков;
3) построение ансамбля из семи классификаторов на основе непересекающихся подмножеств признаков;
4) построение ансамбля из трех классификаторов на основе пересекающихся подмножеств признаков.

Для каждого набора данных был построен как индивидуальный классификатор с использованием подмножества признаков, отобранных с использованием ГА, ранее описанного авторами в [11], так и ансамбль классификаторов.

*Авторы выражают благодарность Мастыкину А.С. (Белорусский государственный медицинский университет, г. Минск, Беларусь) за предоставление данных по ТИА для проведения анализа.

Для проведения экспериментов были выбраны следующие параметры ГА:

– Размер популяции – 100 – 200.
– Максимальное количество генераций – 100.
– Вероятность скрещивания $P_{скр} = 0{,}8$.
– Вероятность мутации $P_{мут} = 0{,}1$.

Результаты экспериментов, полученные для каждого из наборов данных, представлены в табл. 2 – 4. В столбце «Точность классификации» указана точность классификации тестовой выборки. В связи с небольшим количеством признаков, характеризующих объект данных в наборах данных Heart и Wine, ансамбль классификаторов строится с использованием трех или пяти подмножеств признаков.

Таблица 2 – Результаты эксперимента для набора данных Heart

	Количество подмножеств признаков	Точность классификации	Лучшее решение – особь ГА
Классификатор (метод k-ближайших соседей)	1 классификатор	0,754	Все признаки
Классификатор с отбором признаков	1 классификатор	0,829	0,0,1,0,0,0,0,0,0,0,0,1,1
Ансамбль классификаторов (схема 1)	3 классификатора	0,848	0,3,2,0,1,1,3,0,1,0,1,3,1 или 2,1,2,0,1,0,1,0,1,0,2,3,3
	5 классификаторов	0,865	0,1,5,1,1,3,1,0,3,1,2,4,3
Ансамбль классификаторов (схема 2)	3 классификатора	0,865	3,7,2,2,4,3,4,3,7,6,0,5,3

Согласно табл. 2, для набора данных Heart классификатор на отобранном подмножестве наиболее информативных признаков улучшает результаты классификации тестовой выборки с 75,4% до 82,9%. Наилучшие результаты классификации дают ансамбль из пяти классификаторов с непересекающимися подмножествами признаков и ансамбль из трех классификаторов с пересекающимися подмножествами признаков.

Таблица 3 – Результаты эксперимента для набора данных Wine

	Количество подмножеств признаков	Точность классификации	Лучшее решение – особь ГА
Классификатор (метод k-ближайших соседей)	1 классификатор	0,95	Все признаки
Классификатор с отбором признаков	1 классификатор	0,994	1,1,0,0,1,0,1,1,0,1,1,0,1
Ансамбль классификаторов (схема 1)	3 классификатора	0,994	3,1,1,1,2,3,3,0,2,2,3,0,3
	5 классификаторов	0,994	2,0,1,1,2,1,2,2,1,4,4,1,4
Ансамбль классификаторов (схема 2)	3 классификатора	0,994	6,6,1,0,5,5,7,0,2,6,6,1,7

Как видно из табл. 3, для набора данных Wine построенный классификатор на отобранном подмножестве наиболее информативных признаков улучшает результаты классификации тестовой выборки с 75,4% до 82,9%. Точность классификации ансамбля из трех и пяти классификаторов с непересекающимися подмножествами признаков и ансамбля из трех классификаторов с пересекающимися подмножествами признаков не лучше, чем точность отдельного классификатора с отобранным подмножеством признаков. Это можно объяснить тем, что почти все признаки набора данных Wine информативны, что подтверждается высокой точностью классификации тестовой выборки с использованием одного классификатора и всех признаков. Следовательно, построение ансамбля классификаторов путем разбиения множества признаков на несколько подмножеств не улучшает точности классификации и не является необходимым в этом случае.

Таблица 4 – Результаты эксперимента для набора данных ТИА

	Количество подмножеств признаков	Точность классификации
Классификатор (метод k-ближайших соседей)	1 классификатор	0,604
Классификатор с отбором признаков	1 классификатор	0,802
Ансамбль классификаторов (схема 1)	3 классификатора	0,852
	5 классификаторов	0,861
	7 классификаторов	0,792
Ансамбль классификаторов (схема 2)	3 классификатора	0,871

Согласно табл. 4, для набора данных ТИА классификатор на отобранном подмножестве наиболее информативных признаков существенно улучшает результаты классификации тестовой выборки с 60,4% до 80,2%. Наилучшие результаты классификации дают ансамбль из пяти классификаторов с непересекающимися подмножествами признаков (86,1%) и ансамбль из трех классификаторов с пересекающимися подмножествами признаков (87,1%).

Как следует из результатов вычислительных экспериментов с тремя наборами данных, предложенный в настоящей работе подход к построению ансамбля классификаторов обеспечивает получение более высокой точности классификации объектов, характеризующихся большим количеством признаков. Мы надеемся, что это будет подтверждено дальнейшими экспериментами с более широкой номенклатурой тестовых и реальных наборов данных. Как следует из анализа результатов экспериментов, использование в классификаторе всех признаков, включающих как информативные, так и избыточные, дает наихудшие результаты классификации. Использование классификатора, построенного на отборе только одного подмножества информативных признаков, может привести к игнорированию хороших альтернативных решений, которые могут стать составной частью ансамбля классификаторов и в комплексе обеспечить более высокую точность классификации.

Заключение

В представленной работе описан подход к построению ансамбля классификаторов на основе применения модифицированного ГА. Отличительной чертой предложенного подхода является представление задачи построения ансамбля классификаторов

как задачи оптимизации разбиения исходного множества признаков на подмножества, определяющие отдельные классификаторы ансамбля. Применение ГА в качестве инструмента решения оптимизационной задачи позволяет в автоматическом режиме находить такие комбинации классификаторов, которые обеспечивают максимум точности классификации объектов данных ансамблем. Причем в процессе оптимизации происходит определение параметров как отдельных классификаторов, так и их ансамбля. Выполнено тестирование предложенного подхода на нескольких наборах данных, что показало более высокую точность классификации с использованием ансамбля классификаторов, чем с использованием отдельных классификаторов. Дальнейшим направлением исследований является решение задачи построения ансамбля классификаторов с возможностью определения типа для каждого индивидуального классификатора, который будет кодироваться в расширенной хромосоме ГА.

Литература

1. Multiple Classifier Systems / J. Kittler & F. Roli (editors) // Proc. of 2nd International Workshop, MCS2001, (Cambridge, UK, 2-4 July 2001) / Lecture Notes in Computer Science. – Vol. 2096. – Springer-Verlag, Berlin.
2. Vishwath P. Fusion of multiple approximate nearest neighbor classifier for fast and efficient classification / P. Vishwath, M.N. Murty, C. Bhatnagar, // Information fusion. – 2004. – Vol. 5. – P. 239-250.
3. Quinlan J.R. Bagging, boosting and C4.5 / J.R. Quinlan // Proceedings of AAA/IAAI. – 1996. – Vol. 1. – P. 725-730.
4. Tumer K. Decimated input ensembles for improved generalization / K. Tumer, N.C. Oza // Proceedings of the International Joint Conference on Neural Networks. – Washington, DC. – 1999.
5. Bay S.D. Nearest neighbor classifiers from multiple feature subsets / S.D. Bay // Intelligent data analysis. – 1999. – Vol. 3. – P. 191-209.
6. Bryll R. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets / R. Bryll, R. Gutierrez-Osuna, F. Quek // Pattern Recognition. – 2003. – Vol. 36. – P. 1291-1302.
7. Siedlecki W. A note on genetic algorithms for large scale feature selection / W. Siedlecki, J. Sklansky // Pattern Recognition Letters. – 1989. – Vol. 10, № 5. – P. 335-347.
8. Dimensionality reduction using genetic algorithms / Raymer M.L. [et al.] // IEEE Transactions on Evolutionary Computation. – 2000. – Vol. 4, № 2. – P. 164-171.
9. Kuncheva L.I. Nearest neighbor classifier: simultaneous editing and feature selection / L.I. Kuncheva, L.C. Jain // Pattern Recognition Letters. – 1999. – Vol. 20. – P. 1149-1156.
10. Cover T.M. Nearest neighbor pattern classification / T.M. Cover, P.E. Hart // IEEE Transactions on Information Theory. – 1967. – Vol. 13, № 1. – P. 21-27.
11. Новоселова Н.А. Эволюционный подход к выделению информативных признаков в задачах анализа медицинских данных / Н.А. Новоселова, И.Э. Том, А.С. Мастыкин // Искусственный интеллект. – 2008. – № 3. – С. 105-112.

Н.А. Новосьолова, І.Э. Том

Підхід до побудови ансамблю класифікаторів з використанням генетичного алгоритму

У статті розглядається новий еволюційний підхід до побудови ансамблю класифікаторів. Запропонований підхід розроблений на основі генетичного алгоритму з модифікованою схемою реалізації. У процесі оптимізації відбувається визначення параметрів як окремих класифікаторів, так і всього ансамблю. З використанням підходу виконана побудова ансамблю класифікаторів на декількох наборах даних з архіву даних по машинному навчанню й на одному реальному наборі медичних даних. Порівняльне тестування показало переваги використання запропонованого підходу при роботі з багатовимірними даними, що характеризуються більшою кількістю ознак.

N.A. Novoselova, I.E. Tom

Design of Classifier Ensemble by Genetic Algorithm

The paper proposes a new evolutionary approach to classifier ensemble design. The proposed approach is developed on the basis of genetic algorithm with modified realization scheme as applied to the optimization of feature set decomposition into the subsets, which define the individual ensemble's classifiers and provide the high classification accuracy. During optimization both individual classifiers' parameters and the ensemble parameters are defined. With the approach a few ensembles were designed for several datasets from machine learning database and for one real medical dataset. The comparative testing shows the advantages of the proposed approach for multivariate data analysis with great number of features.

Статья поступила в редакцию 15.06.2009.